Making Inferences With Indirect Measurements

Todd Graves and Michael Hamada

Statistical Sciences

Los Alamos National Laboratory

03.14.04 1800

Abstract

This paper considers the characterization of the distribution of part quality for parts produced by a manufacturing process when only indirect quality measurements are available. The proposed method uses the relationship between the indirect and direct measurements which is not completely deterministic as well as a key property of the assumed distribution. The proposed method provides tolerance bounds of the part quality distribution. A diagnostic is also proposed to assess the validity of the distribution assumed. The proposed method and diagnostic are demonstrated with an illustrative data set.

Key Words: scale distribution, scale equivariance, simulation, tolerance interval.

1 Introduction

This paper is concerned with the characterization of the quality of parts produced by a manufacturing process in which only indirect measurements of a quality characteristic are available. Consider a spherical vessel which has an

inner liner. Both the vessel and its liner have defined equators which should coincide. Due to manufacturing variation, however, the equators do not coincide so that what is of interest is characterizing the maximum distances between the equators for the vessels being produced by the manufacturing process. Since maximum distances cannot be negative, we assume that the maximum distances (x's) follow an exponential distribution whose cumulative distribution function is

$$G(x) = 1 - \exp(-x/\mu) \tag{1}$$

for $x \geq 0$, where μ is the mean maximum distance.

Ideally, a distance measurement should be taken at every degree around the vessel's equator to determine the maximum distance. However, due to cost constraints, only one distance measurement is taken at a random angle around the vessel's equator for a sample of vessels. See Figure 1 which displays the liner equator (inner dark line) and the vessel equator (outer light line) that do not coincide. The 0 and 180 degree positions are where the two equators cross. In Figure 1, the maximum distance between the two equators is denoted by x and the distance at randomly chosen θ degrees is denoted by y.

In order to make progress, there needs to be some relationship between the indirect and direct measurements. In this case, we can simply use geometry which yields

$$y = x|\sin(\theta)|. \tag{2}$$

One way to characterize a distribution is to use tolerance bounds, i.e., statistical bounds that capture a large specified proportion of the distribution.

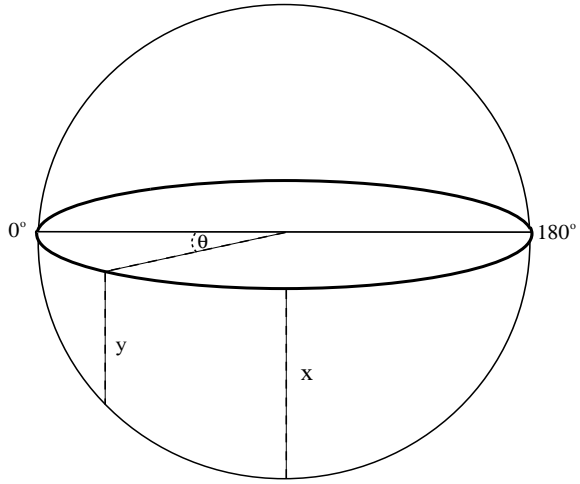

Figure 1: Example diagram.

For the vessel example, what is of interest is providing an upper bound on the maximum vessel-liner distances. Consequently, upper tolerance bounds T are appropriate which are described as follows: with $\beta \times 100\%$ confidence, $\alpha \times 100\%$ of the population of maximum distances is less than T (Hahn and Meeker, 1991).

In this paper, we demonstrate how such inferences about a population can be made with indirect measurements. An outline of the paper is as follows. Section 2 introduces the use of equivariance for the distribution of maximum distances which is assumed to follow a scale distribution. Section 3 consider upper tolerance bounds based on five different estimators and compares them. Section 4 considers diagnostics to be performed on the indirect measurements for assessing the distribution assumed for the direct measurements. Section 5 illustrates the proposed method with a simulated set of data. Section 6

concludes with a discussion.

2 A Solution Using Equivariance

Let $G(\cdot)$ be the cumulative distribution function of the maximum distances given in (1). Given a random sample of n vessels, and for vessel $i = 1, \ldots, n$, a distance y_i measured at a random angle θ_i, then the problem is to find a statistic $T = T(y_1, \ldots, y_n)$ which satisfies

$$Pr\{G(T) \geq \alpha\} \geq \beta. \tag{3}$$

We will see shortly the advantage of restricting ourselves to tolerance bounds T that have the property

$$T(cy_1, \ldots, cy_n) = cT(y_1, \ldots, y_n) \text{ for } c > 0. \tag{4}$$

This property is called *scale equivariance* by Lehmann (1991). Denote by Pr_μ the probability distribution of the x's. Then

$$\begin{aligned} Pr_\mu\{G(T) \geq \alpha\} &= Pr_\mu\{1 - \exp(-T/\mu) \geq \alpha\} \\ &= Pr_\mu\{T \geq -\mu \log(1-\alpha)\} \\ &= Pr_\mu\{T/\mu \geq -\log(1-\alpha)\} \\ &= Pr_\mu\{T(y_1/\mu, \ldots, y_n/\mu) \geq -\log(1-\alpha)\} \\ &= Pr_1\{T \geq -\log(1-\alpha)\}. \end{aligned} \tag{5}$$

The first three equalities follow from substitution and algebra. The fourth equality uses the scale equivariance in (4). The last equality uses the fact that dividing an exponential random variable by its scale parameter results in a standard exponential random variable, i.e., with $\mu = 1$.

The implication of (5) is that we can now propose a statistic T^* satisfying (4), perform simulations of x and y using (2) to obtain the $1-\beta$ quantile $t^*_{1-\beta,n}$ of its distribution and take T to be

$$T(y_1,\ldots,y_n) = -\log(1-\alpha)T^*(y_1,\ldots,y_n)/t^*_{1-\beta,n} \tag{6}$$

which satisfies (3).

The quantity $t^*_{1-\beta.n}$ is easily obtained as follows:

1. Perform steps (a)-(c) n times to obtain simulated $y_1,\ldots,y_n$

   (a) draw x from a standard exponential distribution

   (b) draw θ from a uniform distribution on 0 to 360 degrees

   (c) compute y from (2).

2. Compute T^* from $y_1,\ldots,y_n$.

3. Generate a large number of T^*'s, say N, and estimate $t^*_{1-\beta,n}$ by the $(1-\beta)N$th largest T^*. The notation $t^*_{1-\beta,n}$ indicates that this quantity depends on n, the sample size of the actual data; consequently, a different simulation has to be done for each different n as needed.

Then, for the actual data $y_1,\ldots,y_n$, T can be evaluated using (6) to obtain a valid upper tolerance bound regardless of the value of μ.

3 Comparing Some Estimators

Next, we consider several choices for T^* and compare them. An obvious first choice for T^* is the sample mean of the y's. However, intuition suggests that

larger y's are more informative about the value of μ than smaller values; consequently, the sample mean may not be the best statistic to use. We considered the following five estimators:

- $T_1^*(y) = \bar{y} = n^{-1} \sum_{i=1}^n y_i$, the sample mean;
- $T_2^*(y) = \max_{i=1}^n y_i$, the maximum of the data;
- $T_3^*(y) = \sum_{i=1}^n i y_{(i)} / \sum_{i=1}^n i$, a weighted average of the order statistics of the y's, weighted by their ranks;
- $T_4^*(y) = \sum_{i=1}^n i^2 y_{(i)} / \sum_{i=1}^n i^2$, a weighted average of the order statistics of the y's, weighted by the squares of their ranks;
- $T_5^*(y) = \sqrt{n^{-1} \sum_{i=1}^n y_i^2}$, the square root of the average of the squares of the data.

The last four estimators weight larger y_i's more in estimating μ.

In order to understand the performance of these estimators, we conducted a simulation study for $\alpha = \beta = 0.9$ and $n = 10$, i.e., for samples of size 10 and 90% upper bounds with 90% confidence. For the study, we simulated 10000 samples of $(y_1, \ldots, y_n)$ to estimate the distribution of $T_i^*(y_1, \ldots, y_n)$ for $i = 1, \ldots, 5$. The T_i upper tolerance bounds are based on their respective T_i^*'s. The simulation results are presented in Table 1, whose first two columns give the $t_{1-\beta}^*$ and the expected value of T, respectively. The actual 90% percentile of the exponential distribution with mean one is $-\log(1 - 0.9) = 2.303$, so that the tolerance upper bounds tend to be almost twice as large. The remaining columns provide pairwise comparisons of the candidate upper bounds: the 0.632 in the T_1 row and T_2 column means that the sample mean

based T_1 generates a smaller upper tolerance bound 63% of the time than does the sample maximum based T_2. Somewhat surprisingly, T_1, the sample mean based upper bound, is the best even for sample sizes as small as ten. Also, T_3, T_4, and T_5 are competitive with T_1. The sample maximum based T_2 is not a serious competitor.

Table 1: A Comparison of Five Tolerance Bounds ($\alpha = \beta = 0.9$ and $n = 10$)

	$t^*_{1-\beta}$	$E_1(T)$	$< T_1$	$< T_2$	$< T_3$	$< T_4$	$< T_5$
T_1	0.346	4.21		0.632	0.526	0.567	0.535
T_2	1.003	4.89	0.367		0.351	0.352	0.319
T_3	0.506	4.23	0.474	0.649		0.617	0.533
T_4	0.613	4.29	0.433	0.648	0.383		0.444
T_5	0.484	4.37	0.465	0.681	0.467	0.556	

Next, consider calculating an upper tolerance bounds using the distance data given in Table 2. Take $\alpha = 0.95$ and $\beta = 0.90$. That is, we want upper bounds on 90% of the population with 95% confidence. Here we use the statistic based on the sample mean, T_1^*, for sample size $n = 30$. Based on 100,000 simulations, the $t^*_{(1-\beta,n)}$ for $\beta = 0.95$ is 0.4255. The sample mean $\bar{y}$ for the data in Table 2 is 0.0420. Using (6), we calculate the desired upper bound as 0.227.

4 Model Checking

An important analysis step is to assess whether the assumed exponential distribution is reasonable for the data. A simple graphical method, a quantile-quantile or Q-Q plot, can be used as follows. The probability density function of the indirect measurement y given μ can be derived and has the form:

$$f_\mu(y) = \frac{2}{\pi} \int_y^\infty \mu^{-1} \exp(-x/\mu)(x^2 - y^2)^{-1/2} dx. \tag{7}$$

For sample size n, the $(i - 0.5)/n$, $i = 1, \ldots, n$ quantiles can be calculated using (7) to find the quantiles for which the cumulative distribution function under the standard exponential distribution equals $(i - 0.5)/n$, $i = 1, \ldots, n$. More simply, the quantiles can be obtained by simulating a large number of y's as described in the previous section and using the empirical quantiles. Then the quantile-quantile (QQ) plot is generated by plotting the ordered data $y_{(1)}, \ldots, y_{(n)}$ against the corresponding quantiles; they should plot as a straight line if the exponential distribution assumption holds.

For illustration, consider the case when $n = 30$. Table 2 presents the 30 quantiles under the exponential distribution that were generated from 100,000 simulations as described above. Table 2 also displays sorted distance data simulated from an exponential distribution with $\mu = 0.05$. The QQ plot of these data is shown in Figure 2 which shows the characteristic straight line. To demonstrate that the QQ plot can detect a non-exponential distribution, Figure 3 displays the QQ plot for a sample of size 30 from a $Gamma(5, 100)$ distribution; this QQ plot has a bulge in the middle.

Table 2: Distance Data and Quantiles Based on Exponential Distribution ($n = 30$)

Index	Data	Quantile	Index	Data	Quantile
1	0.002	0.004	16	0.024	0.387
2	0.003	0.015	17	0.030	0.435
3	0.004	0.028	18	0.036	0.488
4	0.004	0.043	19	0.042	0.549
5	0.006	0.059	20	0.042	0.614
6	0.008	0.078	21	0.044	0.686
7	0.009	0.098	22	0.055	0.768
8	0.009	0.121	23	0.059	0.861
9	0.011	0.144	24	0.066	0.973
10	0.016	0.171	25	0.072	1.108
11	0.017	0.199	26	0.073	1.271
12	0.020	0.230	27	0.101	1.475
13	0.020	0.265	28	0.129	1.755
14	0.022	0.302	29	0.145	2.180
15	0.022	0.344	30	0.171	3.111

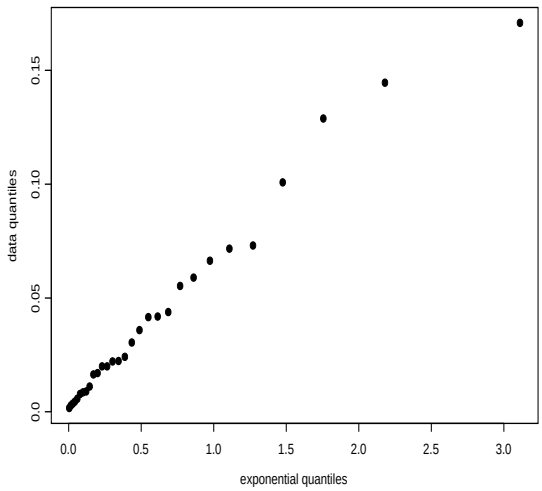

Figure 2: QQ plot when distance data are exponentially distributed ($n = 30$).

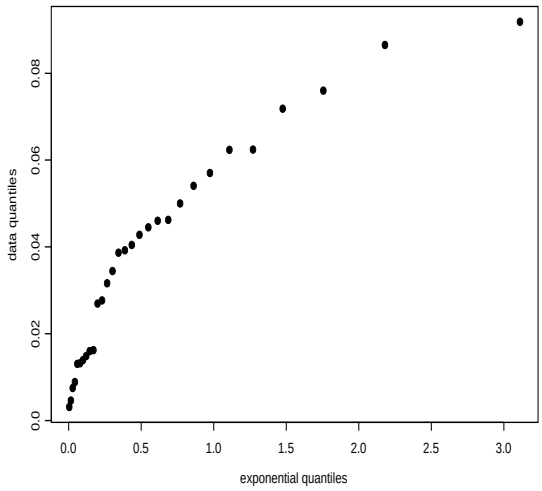

Figure 3: QQ plot when distance data are gamma distributed ($n = 30$).

5 Towards a Generalization

A reviewer suggested that the solution proposed in this paper might be applied more generally. That is, there is a characteristic of interest x whose cumulative distribution function is $G(x|\mu)$, where μ is an unknown parameter. The characteristic x is measured with error; the measurement y is related to x by $y = f(x, \theta)$, where θ is random but whose distribution is known. Then for data $y_1, \ldots, y_n$, one needs to find a statistic $T(y_1, \ldots, y_n)$ that satisfies (3) to obtain an upper tolerance bound. The challenge for a particular problem is to find such a statistic which provides a valid upper tolerance bound no matter what the value of μ.

For the situation considered in this paper, there were a number of factors that made this possible. The error being multiplicative and the unknown μ being a scale parameter allowed estimators with the scale equivariance property to be used. This was key to requiring a single simulation to obtain the value $t^*_{1-\beta,n}$ which provided valid upper tolerance bounds no matter what the value of μ for a given sample size n. Consequently, there may be situations in which the form of the error and the form of the distribution of interest may not allow the generalization of the solution proposed in this paper to be applied.

One such situation, however, that does apply is when the characteristic x is normally distributed with unknown mean μ_x but known variance σ_x^2 and y is the measured x with additive error which is normally distributed with mean zero and known variance σ_m^2. Then estimators with the location equivariance property $((T(y_1 + c, \ldots, y_n + c) = T(y_1, \ldots, y_n) + c)$ can be used to obtain valid upper tolerance bounds.

6 Discussion

In this paper, we showed how indirect measurements in a particular situation could be used to characterize the distribution of the direct measurements for a quality characteristic of interest. We considered how the solution proposed in this paper might be applied more generally. More research on this topic is needed.

Even for the situation considered in this paper, there are a number of issues that could be explored. How large should the sample size n be? If more than one measurement per vessel is taken, how should these data be analyzed? What is the benefit of taking more than one measurement per vessel versus taking one measurement on more vessels? These issues will be considered in a future paper.

Acknowledgements

We thank Dee Won for her encouragement of this work, Cheryll Faust and Patrick Rodriguez for helpful discussions and Jeroen de Mast for insightful comments and suggestions on an earlier version of this paper.

References

Hahn, G.J. and Meeker, W.Q. (1991). *Statistical Intervals: a Guide for Practitioners.* NY: John Wiley & Sons.

Lehmann, E.L. (1991). *Theory of Point Estimation.* NY: Wadsworth & Brooks/Cole.